Note on approximations for the multi-server queue with finite buffer and deterministic service times

Tijms, H.C.

published in
Probability in the Engineering and Informational Sciences
2008

DOI (link to publisher)
10.1017/S0269964808000375

document version
Publisher's PDF, also known as Version of record

Probability in the Engineering and Informational Sciences, **22**, 2008, 653–658. Printed in the U.S.A.
doi:10.1017/S0269964808000375

NOTE ON APPROXIMATIONS FOR THE MULTISERVER QUEUE WITH FINITE BUFFER AND DETERMINISTIC SERVICES

HENK TIJMS
Department of Econometrics and Operations Research
VRIJE University
1081 HV Amsterdam
E-mail: tijms@feweb.vu.nl

This article shows that very accurate accurate approximations to performance measures in the multiserver $M/D/c/c+N$ queue with finite buffer and deterministic service times can be obtained by replacing the deterministic service time by a two-phase process with exponential sojourn times and branching probabilities outside the interval [0, 1].

1. INTRODUCTION

This article extends an earlier article by Tijms and Staats [3] in which approximations were obtained for state probabilities and the waiting times in the single server $M/D/1/N$ queue with finite buffer and deterministic services. These approximation were provided in the form of explicit expressions involving geometric distributions and exponential densities. The underlying idea of the approximation was to replace the deterministic service time by the following two-phase process (see also Fig. 1). The process starts in phase 1. It stays in phase 1 for an exponentially distributed time with mean $1/\gamma$. Upon completion of the sojourn time in phase 1, the process expires with probability r_1 and moves to phase 2 with probability $1-r_1$. The sojourn time in phase 2 is also exponentially distributed with mean $1/\gamma$. Upon completion of the sojourn time in phase 2, the process expires with probability r_2 and returns to phase 1 with probability $1-r_2$. In phase 1, the process starts anew. The idea is to approximate the deterministic service time D by the time it takes until the two-phase process

 0269-9648/08 $25.00

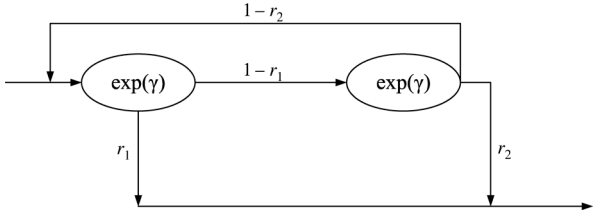

FIGURE 1. Two-phase process.

expires and choose the values of the parameters r_1, r_2, and γ in such a way that the first moments of the sojourn times in the two-phase process matches the first three moments of the service time. This approximation idea is due to Nojo and Watanabe [1]. Independently, a closely related approximation approach was earlier proposed by Van Hoorn and Seelen [4]. However, matching the first three moments is only feasible when values outside the interval $[0, 1]$ are allowed for $r_1$and r_2. Using the Laplace transform

$$f^*(s) = \frac{\gamma r_1 s + \gamma^2(r_1 + r_2 - r_1 r_2)}{s^2 + 2\gamma s + \gamma^2(r_1 + r_2 - r_1 r_2)} \tag{1}$$

of the density of the expiration time in the two-phase process, it is matter of simple algebra to derive that

$$\gamma = \frac{2}{D}, \qquad r_1 = -1, \quad \text{and} \quad r_2 = \frac{5}{4} \tag{2}$$

when the service time is deterministic and equals the constant D. By approximating the deterministic service time by the two-phase process, continuous-time Markov chain analysis can be used for the approximate model by doing calculations with branching probabilities outside the interval $[0, 1]$ as if they were legitimate probabilities.

In the next section we show how approximations can be calculated for various performance measures, including the rejection probability, the delay probability for accepted customers, and the first two moments of both the queue size and the delay in queue of an an accepted customer. These approximations turn out to be surprisingly accurate and improve approximations earlier obtained in Tijms [2]. A question that remains open is how to fit an appropriate probability distribution to the waiting-time distribution of an accepted customer by matching the delay probability and the first two moments of the waiting time. We were not able to find a proper fit to the waiting-time distribution, which is concentrated on the finite interval $[0, \lceil ND/c \rceil]$.

2. MULTISERVER $M/D/C/C + N$ QUEUE

In the multiserver case with c available servers, no explicit expressions can be derived for approximations to the state probabilities and waiting-time probabilities. By replacing the deterministic service time by the two-phase process described in Section 1, we

can use standard continuous-time Markov chain analysis to write down the equilibrium equations for the microstate probabilities

$$p_{nk} = \lim_{n\to\infty} P\{\text{at time } t \text{ there are } n \text{ customers present and } k \text{ of the } \min(n,c) \text{ services in progress are in phase 1}\}$$

for $n = 1, 2, \ldots, N + c$ and $k = 0, 1, \ldots, \min(n, c)$ together with the equilibrium probability $p_0 = \lim_{t\to\infty} P(\text{at time } t \text{ the system is empty})$. The linear equations for the p_{ni} and p_0 must be numerically solved in the multiserver case. (In the single-server case, an explicit expression for the state probabilities was obtained in Tijms and Staats [3].) In practical applications, the computational effort to do so will be not be a bottleneck. Once we have computed these microstate probabilities, we have approximations for the performance measures:

P_{rej} = the long-run fraction of customers who are rejected
P_{delay} = the long-run fraction of accepted customers who have to wait
$E(L_q)$ = the expected value of the queue size L_q in steady state
$\sigma(L_q)$ = the standard deviation of the queue size L_q in steady state.

Denoting by W_q the waiting time in queue of an accepted customer arriving in the steady state, we proceed as follows to compute approximations to the expected value and the standard deviation of W_q.

Obviously,

$$P_{\text{rej}} = \sum_{k=0}^{c} p_{N+c,k} \quad \text{and} \quad P_{\text{delay}} = \sum_{n=c}^{N+c-1} \sum_{k=0}^{c} \frac{p_{nk}}{1 - P_{\text{rej}}}.$$

For the approximating system, define the random variable

$$W_{nk} = \text{the conditional waiting time of an accepted customer who finds upon arrival the system in state } (n,k)$$

and its Laplace-Stieltjes transform

$$f_{n,k}(s) = E[e^{-sW_{nk}}]$$

for $n = c, \ldots, N + c - 1$ and $k = 0, 1, \ldots, c$. Letting

$$f_{c-1,k}(s) = 1 \quad \text{for } k = 0, 1, \ldots, c,$$

we have that the $f_{n,k}(s)$ satisfy nested systems of linear equations. For $n = c$,

$$f_{c,0}(s) = \frac{c\gamma}{c\gamma + s}\left[r_2 f_{c-1,0}(s) + (1 - r_2) f_{c,1}(s)\right],$$

$$f_{c,k}(s) = \frac{c\gamma}{c\gamma + s}\left[\frac{k}{c}\{r_1 f_{c-1,k-1}(s) + (1 - r_1) f_{c,k-1}(s)\} + \frac{c-k}{c}\{r_2 f_{c-1,k}(s) + (1 - r_2) f_{c,k+1}(s)\}\right],$$

$$f_{c,c}(s) = \frac{c\gamma}{c\gamma + s}\left[r_1 f_{c-1,c-1}(s) + (1 - r_1) f_{c,c-1}(s)\right].$$

For each $n = c + 1, \ldots, c + N - 1$,

$$f_{n,0}(s) = \frac{c\gamma}{c\gamma + s}\left[r_2 f_{n-1,1}(s) + (1 - r_2) f_{n,1}(s)\right],$$

$$f_{n,k}(s) = \frac{c\gamma}{c\gamma + s}\left[\frac{k}{c}\{r_1 f_{n-1,k}(s) + (1 - r_1) f_{n,k-1}(s)\} + \frac{c-k}{c}\{r_2 f_{n-1,k+1}(s) + (1 - r_2) f_{n,k+1}(s)\}\right],$$

$$f_{n,c}(s) = \frac{c\gamma}{c\gamma + s}\left[r_1 f_{n-1,c}(s) + (1 - r_1) f_{n,c-1}(s)\right].$$

These equations follow easily from $f_{n,k}(s) = E\left[e^{-sW_{nk}}\right]$ by the memoryless property of the exponential distribution and by a conditioning argument (decompose W_{nk} as the time until the first completion of a service phase plus the remaining waiting time). Taking the first and second derivatives of the $f_{n,k}(s)$ at $s = 0$, we can approximate $E(W_q)$ and $E(W_q^2)$ by

$$E(W_q) = \sum_{n=c}^{c+N-1} \sum_{i=0}^{c} -\frac{p_{ni}}{1 - P_{\text{rej}}} f'_{n,i}(0),$$

$$E(W_q^2) = \sum_{n=c}^{c+N-1} \sum_{i=0}^{c} \frac{p_{ni}}{1 - P_{\text{rej}}} f''_{n,i}(0),$$

respectively. The $f'_{n,i}(0)$ and $f''_{n,i}(0)$ can be computed by solving nested systems of linear equations. It is straightforward to write down these linear equations. We omit the technical details.

In Table 1 we give for several examples the approximate values for the various performance measures together with the simulated values, where, in each case, 20 million customer arrivals were simulated. The system load ρ is defined by $\rho = \lambda D/c$.

TABLE 1. Results for the $M/D/c/c+N$ Queue

	$c=2, N=3,$ $\rho=0.9$		$c=2, N=3,$ $\rho=1.2$		$c=10, N=5,$ $\rho=0.9$		$c=10, N=5,$ $\rho=1.2$	
	App	Sim	App	Sim	App	Sim	App	Sim
P_{rej}	0.087	0.087	0.215	0.215	0.038	0.038	0.183	0.182
P_{delay}	0.705	0.702	0.881	0.879	0.515	0.507	0.883	0.884
$E(L_q)$	0.839	0.844	1.489	1.498	1.006	0.996	2.692	2.720
$\sigma(L_q)$	1.002	1.002	1.076	1.075	1.475	1.474	1.714	1.709
$E(W_q)$	0.510	0.513	0.791	0.796	0.116	0.115	0.275	0.277
$\sigma(W_q)$	0.488	0.486	0.498	0.495	0.155	0.154	0.177	0.172

How about the waiting-time probabilities? The first approach we tried is to apply numerical Laplace inversion to

$$\int_0^\infty e^{-sx} P(W_q > x)\, dx = \frac{1 - E(e^{-sW_q})}{s}$$

with

$$E\left[e^{-sW_q}\right] = 1 - P_{\text{delay}} + \sum_{n=c}^{N+c-1} \sum_{i=0}^{c} p_{ni} f_{n,i}(s).$$

The values of $f_{n,i}(s)$ required in the Laplace inversion algorithm (see Appendix F in Tijms [2]) must be numerically computed by solving systems of linear equations. This approach leads to approximations for $P(W_q > x)$ that are not always satisfactory [e.g., for $c = 1$, $\rho = 0.9$, and $N = 1$, the resulting approximate values of $P(W_q > x)$ are 0.307 and 0.122 for $x = 0.5$ and $x = 1$, whereas the exact values of $P(W_q > x)$ are 0.363 and 0.086, respectively]. This finding is somewhat surprising in view of the fact that the approximations to the state distribution of the number of customers in the system and to the first two moments of W_q are extremely good. In most cases, the approximations for $P(W_q > x)$ obtained by Laplace inversion are practically useful. The quality of numerical results in the example with $c = 2$, $N = 3$, and $\rho = 0.9$ are representative: The conditional probability $P(W_q > x \mid W_q > 0)$ has the values 0.652 (0.668), 0.274 (0.275), 0.134 (0.131), 0.050 (0.040), and 0.011 (0.005) for $x = 0.5$, 1, 1.25, 1.5, and 1.75, respectively, where the values in parentheses indicate the exact values obtained by simulation (50 million customer arrivals). For the example with $c = 2$, $N = 3$, and $\rho = 1.2$, the values 0.799 (0.804), 0.226 (0.233), 0.094 (0.076), 0.025 (0.009), and 0.0028 (0.0001) are obtained for $x = 0.5$, 1, 1.5, 1.75, and 1.95, respectively.

We also tried to fit a simple distribution to the waiting-time distribution of the delayed customers by matching the first two moments. However, using a two-moment match with the beta or the gamma distribution was not very successful. The probability distribution function $P(W_q \le x)$ is concentrated on the finite interval $[0, \lceil ND/c \rceil]$. Also, we tried approximations based on weighted sums of exponential functions on

this finite interval, but this was not successful either. It remains to be further investigated how to approximate the waiting-time distribution through the very accurate approximations to the first two moments of the waiting time.

Acknowledgment

The author is indebted to Koen Staats for his help with the numerical work.

References

1. Nojo, S. & Watanabe, H. (1987). A new stage method getting arbitrary coefficient of variation through two stages. *Transections of the IECIE* 70: 33–36.
2. Tijms, H.C. (2003). *A first course in stochastic models*. Chichester: Wiley.
3. Tijms, H.C. & Staats, K. (2007). Negative probabilities at work in the $M/D/1$ queue. *Probability in the Engineering and Informational Sciences* 21: 69–76.
4. Van Hoorn, M.H. & Seelen, L.P. (1986). Approximations for the $G/G/c$ queue. *Journal of Applied Probability* 23: 484–494.